ED470591 2002-12-00 Some Key Concepts for the Design and Review of Empirical Research. ERIC Digest.

Table of Contents

ERIC Identifier: ED470591
Publication Date: 2002-12-00
Author: Dayton, C. Mitchell
Source: ERIC Clearinghouse on Assessment and Evaluation College Park MD.

Some Key Concepts for the Design and Review of Empirical Research. ERIC Digest.

In the social and health sciences, statistical methods based on probabilistic reasoning are routinely employed in the evaluation of empirical studies. This Digest, intended as an instructional aid for beginning research students and a refresher for researchers in the field, identifies key factors that could play a critical role in determining the credibility that should be given to a specific research study. The need for empirical research, randomization and control, and significance testing are discussed, and seven review criteria are proposed.

NEED FOR EMPIRICAL RESEARCH

In practice, the accumulation of evidence for or against any particular theory involves planned research designs for the collection of empirical data. Several typographies for such designs have been suggested, one of the most popular of which comes from Campbell and Stanley (1963). They are responsible for popularizing the widely cited distinction among pre-experimental, experimental, and quasi-experimental designs and are staunch advocates of the central role of randomized experiments in educational research. In particular, they view the experiment:

...as the only means for settling disputes regarding education practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties (p. 2).

RANDOMIZATION AND CONTROL

The hallmarks of an experiment to Campbell and Stanley, among others are (a) random assignment of cases to comparison groups, (b) control of the implementation of a manipulated treatment variable, and (c) measurement of the outcome with relevant, reliable instruments. Controlled experimentation allows for replication of the conditions of the experiment so that independent researchers can attempt to repeat the results of the experiment. In contrast, non- experimental studies may use convenience samples, comparison groups formed by post-hoc matching and similar procedures.

Campbell and Stanley (1963) provide a framework for evaluating the limitations that various types of research studies pose with respect to inferring a causal link between independent (treatment) and dependent (outcome) variables. They posit a necessary relationship between the validity of an individual research study and the generalization of results from this study to wider populations. They argue that:

Internal validity is the basic minimum without which any experiment is uninterpretable. Did in fact the experimental treatments make a difference in this specific experimental instance? (p. 5).

Typical of potential threats to internal validity are: (1) uncontrolled, extraneous events occurring during the study (called a "history" threat); (2) failure to randomize interviewers or raters across comparison groups (called an "instrumentation" threat); (3) biased or differential selection of cases as occurs when groups are self-selected in a case-control study (call a "selection" threat); and (4) differential loss of cases from comparison groups when there is no pretest to assess the impact of the loss (called an "experimental mortality" threat). Additional threats are discussed in Campbell and Stanley (1963). Note that, in general, control of threats to internal validity allows the research to rule out plausible rival hypotheses concerning differences between comparison groups.

SIGNIFICANCE TESTING

It is unfortunate that the terminology of accepting/rejecting a null hypothesis (that is, determining whether obtained results provide a reason to reject the hypothesis that they are merely a product of chance factors), suggests that these methods are intended to represent final decisions about the reasonableness of a research hypothesis. Indeed this was not the view of R. A. Fisher (1890-1962), who is considered to be the architect of modern statistical design and analysis. In commenting on the distinction between decision theory and statistical inference, Fisher (1959) notes:
An important difference is that Decisions are final while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation but also of revision (p.100).

A test of significance ... is intended to aid the process of learning by observational experience. In what it has to teach each case is unique, though we may judge that our information needs supplementing by further observations of the same, or of a different kind (pp. 100-101).

In the day-to-day work of experimental research in the natural sciences, they [tests of significance] are constantly in use to distinguish real effects of importance to a research programme from such apparent effects as might have appeared in consequence of errors of random sampling, or of uncontrolled variability, of any sort, in the physical or biological material under examination (p.76).

In the final analysis, the accumulation of evidence in favor of a theory may result in general belief that the theory is "correct" in the sense of providing satisfactory explanations for a body of natural phenomena, although this evidence can never reach the level of mathematical certainty. On the other hand, discrepancies between theoretical predictions and natural observations may result in abandonment or revision of the theory. It is in this sense that a scientific theory is falsifiable.

SEVEN CRITERIA FOR EMPIRICAL RESEARCH

(A)Randomization: Ideally, subjects should be randomly selected from the target population and then randomly assigned to treatment conditions. Internal validity (though not external validity) can be attained if available samples are randomly assigned to treatment conditions.
Quasi-experimental designs such as cohort studies require pre-measures and other covariates that allow for statistical adjustment in an attempt to control for history and other threats. Similarly, case-control studies require covariates for adjustment purposes.

However, it should be noted that adjustment for all relevant, non- randomized competing causes in non-experimental studies is an essentially hopeless task. Pedhazur (1997) notes that analysis of covariance (ANCOVA) can be used for increasing precision in experimental studies and for attempting to adjust for initial differences in non-experimental studies. The application of ANCOVA for the first purpose is well founded, and may prove useful in diverse research areas. The applications of ANCOVA for the second purpose, however, is highly questionable because it is fraught with serious flaws (p. 628). Unfortunately, application of ANCOVA in quasi-experimental and nonexperimental research is by and large not valid (p. 654).

(B) Control: Extraneous factors associated with variation in an outcome variable can be controlled by techniques such as selection, stratification, and possibly statistical adjustment or can be randomized. For example, if there are known socio-economic status (SES) differences on a dependent variable, the researcher can: (a) select cases within a relatively narrow range of SES so that its impact becomes negligible or, at least, lessened; (b) stratify experimental cases into SES blocks that can be incorporated into the design and analysis; or (c) obtain a suitable measure of SES and partial out its influence. In experimental settings, the benefit of all of these procedures is to reduce unexplained within-group variation and, thereby, both increase the likelihood of detecting an effect (i.e., increase power) and reduce the uncertainty associated with the magnitude of an effect (i.e., decrease the width of confidence intervals). Alternatively, the research can ignore SES differences, randomly assign cases to groups, and lose the above benefits.

(C) Reliability: It is preferred that outcomes (and covariates) be assessed with relatively little measurement error. Other things being equal, unreliability increases unexplained variation within groups and reduces the power of the analysis. In practice, it may be impractical to assess the reliability of measurement procedures within the scope of a given study, but the selection of measurement instruments should certainly take this factor into consideration. On the other hand, if a study involves observations or ratings by judges, some effort must be undertaken to assure consistency of measurement across raters or judges.

(D) Validity: In selecting a relevant measure for an outcome variable, it is critical that logical inferences can be made from the operationalizations upon which the measure was based to the theoretical constructs relevant to the study. Construct validity refers to the degree to which inferences of this type can legitimately be made.

(E) Implementation of Treatment Variable: An overlooked consideration in many studies is the provision of evidence that the independent variable of interest has actually been applied as intended. Student (1931) described a famous failure of implementation. In 1930 in Scotland the Department of Health conducted the Lanarkshire Milk Experiment to investigate the advantage of giving extra milk to schoolchildren. The experiment, involving 20,000 children, was seriously compromised by some teachers who gave the extra milk to students they considered most needy as opposed to those selected by randomization. The lesson is that there must be some record or documentation supporting the fact that the intended treatment has taken place.

(F) Analysis Issues: Research studies without serious design limitations may nevertheless suffer from inadequate or inappropriate analyses. While there are often alternative analytical approaches that result in equivalent analyses with respect to interpretation of results, it is also the case that inappropriate analysis may limit interpretability. Among issues that arise reasonably often are: (a) failure to utilize an appropriate unit of analysis (e.g., ignoring nesting of students within schools and employing ordinary ANOVA when hierarchical linear modeling would be more appropriate); (b) arriving at models by exploratory procedures but interpreting results as if models were confirmed (e.g., using stepwise multiple regression to "confirm" the importance of predictor variables or using model modification indices in structural equation modeling to alter an initial model to improve fit to data); (c) deriving estimates from complex survey designs without considering design issues (e.g., neither using weighted estimates nor modeling the design when analyzing NAEP data); and (d) ignoring distributional assumptions with parametric procedures such as multiple regression, ANOVA, structural equation modeling, etc. (e.g., ignoring the impact of outliers, extremely skewed distributions of residuals, or lack of homogeneity of variance). There are, of course, many more subtle issues such as the mistaken notion that non-parametric tests for location (e.g., Mann-Whitney U) are insensitive to homogeneity of variance assumptions.

(G) Interpretation Issues: While the use of inferential statistical methods has been a valuable tool in many applied research fields, their use has also led to some unfortunate opportunities to make incorrect or misleading interpretations of results. Recent emphasis on reporting effect sizes may be viewed as valuable, but all too often this takes the form of comparing a computed effect size (e.g., standardized absolute mean difference) with some completelyarbitrary standard (e.g., .5 as indicating a "medium" effect). In fact, a statistically significant outcome for, say, a two-independent-sample t test for means merely suggests that the result is "surprising" when compared to a model of chance variation. The practical interpretation of the observed outcome must be made within the context of the research setting.

REFERENCES

Campbell, D. T. & Stanley, J. C. (1963). Experimental and Quasi-Experimental Designs for Research.. Chicago: Rand McNally.
Fisher, R. A. (1959). Statistical Methods & Scientific Inference. New York: Hafner Publishing.

Pedhazur, E. J. (1997). Multiple Regression in Behavioral Research (Third Edition). Fort Worth: Harcourt Brace. Student (1931). The Lanarkshire milk experiment. Biometrika, 23, 398-404.

This publication was prepared with funding from the Office of Education Research and Improvement, U.S. Department of Education, under contract ED-99-CO-0032. The opinions expressed in this report do not necessarily reflect the positions or policies or the U.S. Department of Education. Permission is granted to copy and distribute this ERIC/AE Digest.

Title: Some Key Concepts for the Design and Review of Empirical Research. ERIC Digest.
Document Type: Information Analyses---ERIC Information Analysis Products (IAPs) (071); Information Analyses---ERIC Digests (Selected) in Full Text (073);
Available From: ERIC Clearinghouse on Assessment and Evaluation, 1129 Shriver Laboratory, University of Maryland, College Park, MD 20742. Tel: 800-464-3742 (Toll Free); Web site: http://ericae.net.
Descriptors: Credibility, Data Analysis, Reliability, Research, Research Reports, Validity
Identifiers: ERIC Digests, Randomization, Statistical Process Control
###

[Return to ERIC Digest Search Page]